DOCUMENT RESUME

ED 278 264 FL 016 403

AUTHOR	Clark, John L. D.; Li, Ying-che
TITLE	Development, Validation, and Dissemination of a Proficiency-Based Test of Speaking Ability in Chinese and an Associated Assessment Model for Other Less Commonly Taught Languages.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
SPONS AGENCY	Department of Education, Washington, DC.
PUB DATE	Dec 86
GRANT	G008402258
NOTE	34p.; For related document, see FL 016 404.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Audiotape Recordings; *Chinese; Comparative Analysis; Correlation; Evaluation Methods; Interviews; *Language Proficiency; *Language Tests; Models; Speech Skills; Test Construction; Test Use; Test Validity; *Uncommonly Taught Languages; *Verbal Tests
IDENTIFIERS	ACTFL ILR Oral Proficiency Guidelines

ABSTRACT

A project to develop and validate a tape-based alternative to the interview-based speaking proficiency test is described. The objective was to produce an economically more viable test for less commonly taught languages, closely modeled on and readily interpretable in terms of the American Council on the Teaching of Foreign Languages/Interagency Language Roundtable (ACTFL/ILR) proficiency guidelines. The process involved development of four test versions for Chinese and comparison of a representative student sampling's pilot test results with results obtained with the ACTFL/ILR interview and rating procedure. Validation results showed a substantial correspondence between scores on the tape-based test and the live interview results, supporting the taped test as an appropriate and effective alternative to the live interview in situations where use of the latter is not financially or administratively feasible. The description of the test development and validation process which makes up most of the document consists chiefly of detailed explanations of the 34 statistical tables and 9 figures which are included and which display the basic data of the study. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED278264

Development, Validation, and Dissemination of a Proficiency-Based Test of Speaking Ability in Chinese and an Associated Assessment Model for Other Less Commonly Taught Languages

Final Project Report for

Grant No. G008402258

U. S. Department of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

John L. D. Clark

Project Director

Ying-che Li

Project Co-Director

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Center for Applied Linguistics
1118 22nd Street, NW
Washington, DC 20037

December 1986

FL016403

BEST COPY AVAILABLE

OVERVIEW OF PROJECT

The developing "proficiency" movement within the U.S. language teaching field--in large part stimulated by the efforts of the American Council on the Teaching of Foreign Languages (ACTFL) and with the assistance of several government language training agencies under the auspices of the Interagency Language Roundtable (ILR)--has placed a premium on the accurate and reliable measurement of functional language skills, especially listening comprehension and speaking, within a real-life language use context. Face-to-face speaking proficiency tests of the ACTFL/ILR type (i.e., "live" interviews conducted by a trained interviewer/rater and scored on the basis of the ACTFL/ILR verbal descriptive guidelines) have been quite widely implemented within the larger-volume languages such as French, Spanish, and German by means of tester-training workshops and associated testing networks. However, for most of the less-commonly-taught language programs in the United States, the economic and organizational realities, at least for the present and near-term future, are such as to preclude the development and deployment of sufficient cadres of trained interviewers and raters to adequately meet the speaking testing needs at issue in the adoption of a "proficiency" approach to language instruction.

The major objective of the reported project was to develop and make available a tape-based, alternative approach to interview-based speaking proficiency testing that would be economically viable for use with languages having relatively low student volumes, but that at the same time would be closely modeled on, and readily interpretable in terms of, the ACTFL/ILR proficiency guidelines. To accomplish this goal, four alternate forms of such a test were developed in Chinese and validated on a representative student population through direct statistical comparison with the regular ACTFL/ILR interview and rating procedure. Validation results showed a substantial correspondence between student scores on the tape-based tests and the results the "live" interview, providing considerable confidence in the value of the taped test as an appropriate and effective alternative to the live interview in situations where use of the the latter was not financially or administratively feasible. The second, closely related objective was to develop an informational handbook describing the overall project, the format and content specifications developed for the tests and the rationales underlying these specifications, and the step-by-step procedures followed in producing, administering, and evaluating these instruments, for use by interested individuals, organizations, or groups considering the development of similar assessment procedures for other less-commonly-taught languages. Since much of the information concerning test development and administration procedures that would ordinarily appear in a final project report is provided in this handbook (included for reference at the end of this report), the following pages will concentrate on an overview description of major project activities and on a procedural and statistical report of the test validation phase of the project. The reader is asked to refer to the handbook section for more detailed discussion of test rationale and test development, as well as to the Chinese test booklets and scripts themselves, which are reproduced in Appendix A.

MAJOR PROJECT ACTIVITIES

Project Working Committe Meeting/Initial Test Planning

The day-to-day work of the project was conducted primarily at the Center

BEST COPY AVAILABLE

for Applied Linguistics (CAL) by the project director, John L. D. Clark, working with other CAL staff and in close coordination with the project co-director, Dr. Ying-che Li (University of Hawaii). Project planning, review of materials, and expert consultation throughout the project period was provided by a Working Committee consisting of, in addition to the project director and co-director:

Dr. Albert Dien (Stanford University)
Dr. Shang-Hsien Ho (University of Hawaii)
Dr. Timothy Light (Ohio State University)
Dr. Eugene Liu (University of Pennsylvania)
Dr. Pardee Lowe (CIA Language School)
Dr. A. Ronald Walton (University of Maryland)

A major planning meeting of the committee was held on August 3-5, 1985, in which both the proposed format and question types for the test were developed, subject to possible modifications on the basis of clinical tryouts of a draft form of the test. The test as initially designed consisted of five separate sections, as follows:

Personal conversation - Student listens to conversational questions in Chinese and responds to each question as it is asked.

Single picture descriptions - Examinee looks at detailed line drawings and answers questions about them.

Picture sequences - Examinee describes a series of events in a narrative fashion, based on a sequence of 3-5 line drawings.

English-cued discourse - Relatively longer discourse is elicited by means of printed English questions to which the examinee replies in Chinese.

Situations - Examinee reads a printed description of a real-life situation in which a specified interlocutor and communicative task are identified. Examinee is to carry out the indicated task.

Draft Test Administration

Over November 1984 - January 1985, a preliminary version of the test based on the above content specifications was administered to a total of 27 students of Chinese at five institutions: Cornell University , Defense Language Institute (Monterey), University of Hawaii, University of Pennsylvania, and Stanford University. Based on this tryout, the overall format and content of the test were generally confirmed, with relatively minor modifications suggested (e.g., some shortening of the English directions; moderate increase in time allotted for student response to certain questions, etc.).

Preparation of Final Test Forms/Validation Administration

On the basis of the information obtained during the trial administration, four separate final forms of the test were developed, each using similar formats and question types but having different topical content. To validate each of the tests, both as (1) constituting essentially interchangeable versions of the test (i.e., producting similar examinee results independently

of the particular form administered) and (2) producing the same scoring result as a regular "live" interview for any given examinee, an administration plan was developed in which each of 32 students took two forms of the project-developed test, as well as a face-to-face interview.

Based on test administrations conducted at Brigham Young University and at the University of Hawaii in spring 1986, the resulting student response tapes (3 per student--one for each of two of the project-developed tests and one of the live interview) were independently scored by each of two certified ACTFL interviewer/raters, a procedure which allowed for the determination of both the inter-rater reliability of the project tests and the extent of correlation between the project tests and the scores assigned on the basis of the face-to-face interview. These results are described and discussed in the following section.

Validation Study and Results

As indicated above, the test validation paradigm used in this study consisted of the administration of a highly face-valid criterion instrument (face-to-face interview using ACTFL-trained interviewer/raters) and two forms of the experimental semi-direct test to each of 32 native English-speaking learners of Chinese, consisting of 16 students of Chinese at the University of Hawaii and 16 students at Brigham Young University. At both institutions, participating students were drawn from among a volunteer group expressing interest in the project, with final selection made (on the basis of instructors' general familiarity with their speaking performance) so as to provide a wide and, to the extent possible, rectangular distribution of proficiency levels. Each participating student received a small honorarium of $10 for the approximately 1-1/2 total hours of testing involved. The Chinese specialists responsible for administering and rating the live interviews as well as for listening to and rating the student response tapes for each of the semi-direct tests were Dr. Richard Chi of Brigham Young University and Dr. Shang-Hsien Ho of the University of Hawaii, both ACTFL-certified interviewer/raters in Chinese.

For all students at both BYU and Hawaii, the live interview was administered first, followed by two forms (out of the total of four forms) of the tape-administered semi-direct test. Designation of the two particular forms to be administered to a given student, as well as the order in which the forms were administered, was on the basis of a Latin square design which served to control for possible sequence effects in test form administration. To rule out the possibility that prior knowledge of a student's language background or general level of ability would influence the face-to-face interviewing procedure and/or rating assigned, Dr. Chi conducted each of the interviews of the University of Hawaii students and Dr. Ho carried out each of the live interviews at BYU.

Except in 2-3 instances in which scheduling diffficulties mandated the testing of a given student over a two-day period, all tests (live interview plus two taped-test forms) were administered within a single day. Informal conversation with the students following test administration indicated that they did not consider the amount of testing (approximately 15-35 minutes for the live interview and 35 minutes each for the taped tests) unduly onerous or fatiguing. In addition to the cassette recordings of the taped test, audio recordings of the face-to-face interview were also made, using lavalier

microphones which picked up well both the interviewee's and tester's voice. During the interview, the tester was free to make brief notes as desired, but the official scoring was in all instances based on a later re-listening to the interview tape recording by the original interviewer and, as a check on interrater reliability, by the second rater.

On completion of test administration, the cassette tape recordings of the 32 live interviews (16 students at each of the two institutions) were assigned arbitrary code numbers and were randomly sequenced for independent rating by both Dr. Chi and Dr. Ho. Similarly, all 64 cases of student responses to the semi-direct tests (8 for each of Forms A, B, C, and D; for each of the two institutions) were assigned an identification code and randomly sequenced for rating. The entire rating process for all of the face-to-face interviews and semi-direct tests occupied each of the raters on a partial-time basis over an approximately six-week period. In general, during this period, each of the raters would listen to and evaluate a series of several of the semi-direct tests, interspersed with the rating of a portion of the interview tapes. This served to break the monotony of scoring the large number of semi-direct tests involved, and also provided an opportunity for the rater to "re-anchor" himself in the traditional interview/interview scoring process from time to time in the course of rating the semi-direct tests.

For both the live-interview recordings and semi-direct tests, the rater was asked to "fast forward" the tape past the spoken name of the student (given at the beginning of the live interview and as the first question of the semi-direct test) and to evaluate and record the rating results only with reference to the indicated code number. Discussions with the two raters in the course of the rating process indicated that the tapes were being evaluated on an essentially anonymous basis and that, as a result of the large total number of tapes and the random presentation sequence, there was little or no recognition of individual students or recollection of their performance in the live interview.

Rating of both the live interview and the tape-based semi-direct tests was done on a 13-point scale combining both ACTFL and ILR rating scales as follows:

ACTFL/ILR Level	Coded As:
Novice-Low	01
Novice-Mid	02
Novice-High	03
Intermediate-Low	04
Intermediate-Mid	05
Intermediate-High	06
Advanced	07
Advanced-Plus	08
Level 3	09
Level 3+	10
Level 4	11
Level 4+	12
Level 5	13

The several tables below provide descriptive statistics, interrater

reliabilities, and test-retest data obtained in the study. Table 1 shows the range, mean score, standard distribution, and other basic statistics for the ratings assigned by each of the two raters, Dr. Chi (hereafter, Rater 1) and Dr. Ho (Rater 2), to student performances on each of the semi-direct test forms and on the live interview.

Table 1

Descriptive Statistics for Scoring Levels Assigned, Taped and Live Tests

Test Form	Rater 1	Rater 2
	RANGE	
A (N=16)	4-11	4-11
B (N=15)	4-11	4-11
C (N=16)	5-11	4-10
D (N=16)	5-11	4-10
Interview (N=32)	4-11	4-12
	MEAN	
A	8.0	6.9
B	7.8	6.9
C	7.6	6.6
D	7.3	6.5
Interview	7.7	7.3
	MEDIAN/MODE	
A	8/8	7/7
B	8/8	7/7
C	8/5,8 (bimodal)	7/7
D	8/8	6.5/4
Interview	8/8	7/7
	STANDARD DEVIATION	

A	2.2	2.1
B	1.9	1.9
C	2.0	1.8
D	1.8	2.0
Interview	1.9	2.0

Interrater reliabilities (Pearson product-moment correlations) between the ratings assigned by Rater 1 and those assigned by Rater 2 for each of the semi-direct test forms and for the live interview are shown in Table 2 below.

Table 2

Interrater Reliabilities

Test Form	Correlation
A	.89
B	.96
C	.93
D	.91
Interview	.88

Test-retest reliabilities for the same student taking two different test forms, with the same rater scoring both forms, are shown in Table 3.

Table 3

Test-Retest Reliabilities (Same Rater)

Tests Taken By Student	Rater 1	Rater 2
Forms A and B	.95	.99
Forms C and D	.95	.93

Table 4 shows test-retest reliabilities for students taking two different test forms, each form scored by a different rater.

Table 4

Test-Retest Reliabilities (Different Forms and Raters)

Rater/Form Combination	Correlation
Rater 1/Form A - Rater 2/Form B	.90
Rater 1/Form B - Rater 2/Form A	.94
Rater 1/Form C - Rater 2/Form D	.91
Rater 1/Form D - Rater 2/Form C	.91

Correlations of semi-direct test scores with the live face-to-face interview are given in Table 5 below.

Table 5

Correlations with Live Interview

Rater/Form	Inter. as Scored by Rater 1	Inter. as Scored by Rater 2
Rater 1/Form A	.98	.86
Rater 1/Form B	.97	.91
Rater 1/Form C	.96	.90
Rater 1/Form D	.97	.89
Rater 2/Form A	.90	.98
Rater 2/Form B	.93	.97
Rater 2/Form C	.92	.92
Rater 2/Form D	.91	.92

Interrater reliability of the live interview scoring was .88. (Test-retest reliability information for the live interview is not available, since all interview scoring was based on both raters listening to a single tape-recorded interview for any given student.)

As a general summary of the statistical information above, it may be stated that all four forms of the experimental semi-direct test reveal high

interrater reliabilities, with Pearson product-moment correlations uniformly at around the .90 level or higher. Test-retest reliabilities are also in the .90 and higher range under the most "severe" conditions (i.e., different raters rating two different forms) and are even higher (mid-.90s) for test-retest comparisons involving the same rater. Test validity coefficients (correlations of the semi-direct tests against the live interview as an external criterion) are also very high, ranging from .86 to .98, with a mean value of .93 across 16 different test form/rater/interview rater combinations.

These correlation results appear to indicate that there is a strong and consistent linear relationship among sets of assigned scores on the four semi-direct tests for a given group of students, from the three basic standpoints of interrater reliability, test-retest ("alternate form") reliability, and correlation with an external more highly face- and content-valid criterion measure. However, in addition to reliability coefficients per se, a second aspect of test performance which requires examination is the extent to which the absolute values of the assigned ratings remain the same across different raters and test forms. Alternatively stated, it is necessary to determine the extent to which:

(a) for any given semi-direct test form, students will receive similar scores regardless of the particular rater evaluating that test;

(b) students who receive a given score on one form of the test will receive the same score on each of the other test forms, whether these are rated by the original rater or some other rater;

(c) whether a student rated at a given level on the basis of the semi-direct test will receive the same rating on a live face-to-face interview.

Tables 6-33 show the two-way crosstabulations of raters, semi-direct test scores, and interview scores that address these three questions.

Interrater reliability. As indicated in Tables 6-9, there appears to be a clear pattern of at least slight generosity in rating on the part of Rater 1 by comparison to the scores assigned by Rater 2--a tendency which is in evidence across all four forms of the test. For the most part, the magnitude of the discrepancy is a single point difference (for example, "Intermediate-Mid" vs. "Intermediate-High,")--a degree of variation that would be characterized as within a "plus" point on the regular ACTFL/ILR scale. However, in a few instances, the difference is a full level, and in one case (Table 6), two full levels, an apparent "one-time" anomaly considering the rating pattern as a whole. For comparision purposes, crosstabs for interrater reliability for the live interview scores (Table 34) show a similar clear pattern of proportionately higher scores on the part of Rater 1.

Test-retest reliability. Cross-tabs for different test forms as scored by the same rater are shown in Tables 10-13. Overall, there appears to be a strong absolute-value correspondence between the scores assigned on alternate forms of the semi-direct test as these are being scored by the same rater. Form A vs. Form B as scored by Rater 1 shows only occasional "plus-point" variations which are not consistently in either direction; for Form C vs. Form D, four of the 16 data pairs show plus-point generosity in favor of Form C, with one full-level difference in the same direction observed. Rater 2 gives virtually identical

Table 6

X-Axis: Rater 1 - Form A

Y-Axis: Rater 2 - Form B

		(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low	(4)	2				1			
Inter-Mid	(5)			2					
Inter-High	(6)								
Adv	(7)				1	4	2		
Adv-Plus	(8)								
Level 3	(9)								1
Level 3+	(10)							1	1
Level 4	(11)								1

Table 7

X-Axis: Rater 1 - Form B

Y-Axis: Rater 2 - Form B

		(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low	(4)	1	1						
Inter-Mid	(5)			1	1				
Inter-High	(6)			1					
Adv	(7)					6	1		
Adv-Plus	(8)								
Level 3	(9)							1	
Level 3+	(10)								1
Level 4	(11)								1

Table 8

X-Axis: Rater 1 - Form C

Y-Axis: Rater 2 - Form C

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		2	1					
Inter-Mid (5)		2						
Inter-High (6)			1		1			
Adv (7)					3	2		
Adv-Plus (8)							1	
Level 3 (9)						1	1	
Level 3+ (10)								1
Level 4 (11)								

Table 9

X-Axis: Rater 1 - Form D

Y-Axis: Rater 2 - Form D

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		4						
Inter-Mid (5)		1	1					
Inter-High (6)					2			
Adv (7)					2			
Adv-Plus (8)					3			
Level 3 (9)						1		1
Level 3+ (10)						1		
Level 4 (11)								

Table 10

X-Axis: Rater 1 - Form A

Y-Axis: Rater 1 - Form B

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	1							
Inter-Mid (5)	1							
Inter-High (6)			2					
Adv (7)					1			
Adv-Plus (8)				1	4	1		
Level 3 (9)						1		
Level 3+ (10)								1
Level 4 (11)							1	1

Table 11

X-Axis: Rater 1 - Form C

Y-Axis: Rater 1 - Form D

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)		4	1					
Inter-High (6)			1					
Adv (7)								
Adv-Plus (8)					4	2	1	
Level 3 (9)						1	1	
Level 3+ (10)								
Level 4 (11)								1

Table 12

X-Axis: Rater 2 - Form A

Y-Axis: Rater 2 - Form B

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2							
Inter-Mid (5)	1	1						
Inter-High (6)		1						
Adv (7)				7				
Adv-Plus (8)								
Level 3 (9)						1		
Level 3+ (10)							1	
Level 4 (11)								1

Table 13

X-Axis: Rater 2 - Form C

Y-Axis: Rater 2 - Form D

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2	2						
Inter-Mid (5)	1		1					
Inter-High (6)			1	1				
Adv (7)				2				
Adv-Plus (8)				2	1			
Level 3 (9)						1	1	
Level 3+ (10)						1		
Level 4 (11)								

Table 14

X-Axis: Rater 1 - Form A

Y-Axis: Rater 2 - Form B

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2							
Inter-Mid (5)			1		1			
Inter-High (6)			1					
Adv (7)				1	4	2		
Adv-Plus (8)								
Level 3 (9)								1
Level 3+ (10)							1	
Level 4 (11)								1

Table 15

X-Axis: Rater 1 - Form B

Y-Axis: Rater 2 - Form A

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	1	1		1				
Inter-Mid (5)			2					
Inter-High (6)								
Adv (7)					6	1		
Adv-Plus (8)								
Level 3 (9)							1	
Level 3+ (10)								1
Level 4 (11)								1

Table 16

X-Axis: Rater 1 - Form C

Y-Axis: Rater 2 - Form D

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		4						
Inter-Mid (5)			2					
Inter-High (6)					1	1		
Adv (7)					2			
Adv-Plus (8)					1	1	1	
Level 3 (9)							1	1
Level 3+ (10)						1		
Level 4 (11)								

Table 17

X-Axis: Rater 1 - Form D

Y-Axis: Rater 2 - Form C

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		2	1					
Inter-Mid (5)		2						
Inter-High (6)		1			1			
Adv (7)					5			
Adv-Plus (8)					1			
Level 3 (9)						2		
Level 3+ (10)								1
Level 4 (11)								

Table 18

X-Axis: Rater 1 - Form A .

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2							
Inter-Mid (5)								
Inter-High (6)			1					
Adv (7)			1					
Adv-Plus (8)				1	5	1		
Level 3 (9)						1		
Level 3+ (10)								
Level 4 (11)							1	3

Table 19

X-Axis: Rater 1 - Form B

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	1	1						
Inter-Mid (5)								
Inter-High (6)			1					
Adv (7)								
Adv-Plus (8)				1	6			
Level 3 (9)						1		
Level 3+ (10)								
Level 4 (11)							1	2

Table 20

X-Axis: Rater 1 - Form C

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		1						
Inter-Mid (5)		3						
Inter-High (6)			2					
Adv (7)					1			
Adv-Plus (8)					3	2		
Level 3 (9)						1	1	
Level 3+ (10)					1			1
Level 4 (11)								

Table 21

X-Axis: Rater 1 - Form D

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)		4						
Inter-High (6)		1	1					
Adv (7)					1			
Adv-Plus (8)					6			
Level 3 (9)						2		
Level 3+ (10)								1
Level 4 (11)								

Table 22

X-Axis: Rater 2 - Form A

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	3							
Inter-High (6)		2						
Adv (7)				5				
Adv-Plus (8)				2				
Level 3 (9)								
Level 3+ (10)						1		
Level 4 (11)							2	

Table 23

X-Axis: Rater 2 - Form B

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	2	1						
Inter-High (6)		1	1					
Adv (7)				5				
Adv-Plus (8)				2				
Level 3 (9)								
Level 3+ (10)						1		
Level 4 (11)							1	
Level 4+ (12)								1

Table 24

X-Axis: Rater 2 - Form C

Y-Axis: Rater 2 Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		1						
Inter-Mid (5)	2							
Inter-High (6)	1	1	1					
Adv (7)			1	3				
Adv-Plus (8)				2	1			
Level 3 (9)						1		
Level 3+ (10)						1	1	
Level 4 (11)								

Table 25

X-Axis: Rater 2 - Form D

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	1	1						
Inter-Mid (5)	1							
Inter-High (6)	2	1						
Adv (7)			2	2				
Adv-Plus (8)					3			
Level 3 (9)							1	
Level 3+ (10)						2		
Level 4 (11)								

Table 26

X-Axis: Rater 1 - Form A

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	2				1			
Inter-High (6)			2					
Adv (7)					4	1		
Adv-Plus (8)				1		1		
Level 3 (9)								
Level 3+ (10)								1
Level 4 (11)							1	1

Table 27

X-Axis: Rater 1 - Form B

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	1	1		1	1			
Inter-High (6)			2					
Adv (7)					4	1		
Adv-Plus (8)					2			
Level 3 (9)								
Level 3+ (10)							1	
Level 4 (11)								1
Level 4+ (12)								1

Table 28

X-Axis: Rater 1 - Form C

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		1						
Inter-Mid (5)			1					
Inter-High (6)		3	1					
Adv (7)					3	1		
Adv-Plus (8)					1	1	1	
Level 3 (9)						1		
Level 3+ (10)							1	1
Level 4 (11)								

Table 29

X-Axis: Rater 1 - Form D

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		1						
Inter-Mid (5)		1	1					
Inter-High (6)		3						
Adv (7)					4			
Adv-Plus (8)					3			
Level 3 (9)						1		
Level 3+ (10)						1		1
Level 4 (11)								

Table 30

X-Axis: Rater 2 - Form A

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2							
Inter-Mid (5)								
Inter-High (6)		1						
Adv (7)		1						
Adv-Plus (8)	1			6				
Level 3 (9)				1				
Level 3+ (10)								
Level 4 (11)						1	2	1

Table 31

X-Axis: Rater 2 - Form B

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)	2							
Inter-Mid (5)								
Inter-High (6)		1						
Adv (7)			1					
Adv-Plus (8)		1		6				
Level 3 (9)				1				
Level 3+ (10)								
Level 4 (11)						1	1	1

Table 32

X-Axis: Rater 2 - Form C

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	2	2						
Inter-High (6)	1		1					
Adv (7)				1				
Adv-Plus (8)			1	4	1			
Level 3 (9)						2		
Level 3+ (10)							1	
Level 4 (11)								

Table 33

X-Axis: Rater 2 - Form D

Y-Axis: Rater 1 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)								
Inter-Mid (5)	4							
Inter-High (6)		2						
Adv (7)					1			
Adv-Plus (8)			2	2	2			
Level 3 (9)						1	1	
Level 3+ (10)						1		
Level 4 (11)								

Table 34

X-Axis: Rater 1 - Interview Score

Y-Axis: Rater 2 - Interview Score

	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Inter-Low (4)		1						
Inter-Mid (5)	2	1	1		1			
Inter-High (6)		2	2	1				
Adv (7)					8	1		
Adv-Plus (8)				1	4			
Level 3 (9)						1		
Level 3+ (10)						1	1	1
Level 4 (11)								2
Level 4+ (12)								1

scores for Forms A and B, with only two plus-point discrepancies in total. For Forms C and D, all discrepancies are within a "plus" point, but in a greater total number of instances (8 of 16); there is no apparent pattern of generosity for one or another of these test forms. As a general summary, quite similar test-retest scores were observed for both raters in the two form-to-form comparisions available (A/B and C/D), with a large number of identical scores in the two instances and with practically no discrepancy greater than a "plus" point.

As would be expected, there is somewhat greater inter-form variation in assigned scores when different raters, as well as different forms, are involved. As indicated rather clearly in Tables 14-17, scores assigned by Rater 1 are almost always higher than those assigned by Rater 2, a tendency which is in evidence across all of the test form comparisons involved (A/B, B/A, C/D, D/C). In addition, the magnitude of the discrepancy reflects a full-level difference in several instances.

Criterion validity. Tables 18-33 show correspondences between scores assigned to students on the semi-direct tests by comparision to those given on the basis of the live interview. These results appear very similar to those obtained for test-retest comparisions involving alternate forms of the semi-direct test, as discussed immediately above. When both the semi-direct test and the interview are evaluated by the same rater (Tables 18-21 for Rater 1; Tables 22-25 for Rater 2), the obtained pairs of scores are either identical or (except for an occasional "full-level" discrepancy on the part of Rater 2) differ by no more than a "plus" point in either direction. However, when different raters are involved, either in evaluating the semi-direct test or the interview, the scoring differences are much more appreciable, with Rater 1 clearly more generous than his colleague, regardless of whether Rater 1 is evaluating the student on the basis of the semi-direct test (Tables 26-29) or the interview (Tables 30-33).

In summary of the above, it may be suggested that, to the extent warranted by the two-rater, four-test-form comparisions available in this study, it is possible to obtain a high level of congruence of the absolute values of assigned ratings (as well as very high product-moment correlations) between both the various forms of the semi-direct test and between the semi-direct test and live interview scores when the tests/interviews are being evaluated by a single rater. When two different raters are involved, either scoring the same or different test forms or the live interview vs. the taped test, an appreciably higher incidence of differences in the absolute values of the scores is shown, even though the linear correlations themselves remain high.

With regard to the practical applications of the semi-direct test, and pending further examination of the scoring reliability of both the semi-direct test and the live interview in a number of other contexts involving a variety of different raters and examinee populations, the following conclusions may be tentatively made:

(1) Holding the rater constant, different forms of the semi-direct test provide largely equivalent scoring results on a test-retest basis.

(2) The semi-direct test forms developed in this study provide scoring results that are largely equivalent to those obtained in the live interview,

again holding the rater constant for both types of test.

(3) Some discrepancy may be anticipated in the scoring results for one form of the semi-direct test vs. another form, or for semi-direct test vs. live interview results, when different raters are used to provide these data. The observed discrepancies do not appear to be attributable to the format or other characteristics of the semi-direct test as such, but also occur with about the same magnitude and effect in the scoring of the live interview by two separate raters. In both types of testing, close attention to the nature and effectiveness of the rater training process as related to the participants' developed ability to assign identical scores to a given candidate performance would seem to be in order.

STUDENT FEEDBACK ON SEMI-DIRECT TESTING

Feedback information from the participating students concerning various aspects of their experience with and opinions about the semi-direct testing procedure were elicited by means of a short questionnaire (Appendix B) which was administered after both the live and semi-direct testing had been completed. Twenty-seven of the total of 32 participants (84%) returned a completed questionnaire, the results of which are summarized below.

The first two questions asked for a student comparison between the live interview and semi-direct formats in terms of the extent to which each of these testing approaches had succeeded in eliciting the highest level of language performance of which they were capable. The two questions read as follows:

"Over the course of the live interview, do you feel that your maximum level of speaking ability in Chinese was adequately probed by the tester?"

"Over the course of the taped test, do you feel that the descriptions, narrative situations, and other types of questions in the test were adequate to probe your maximum level of speaking proficiency in Chinese?"

The generally comparable relative percentages of "yes" and "no" responses for each of these questions (Figures 1 and 2) suggest that the examinees for the most part found no difference between the live and taped formats with respect to the depth and thoroughness with which their maximum speaking performance had been elicited. A second pair of questions asked for a similar comparision of the overall "fairness" of the two testing approaches:

"In the live interview, were there any questions asked or speaking situations required which you felt were in any way 'unfair'?"

"In the taped test, were there any picture/descriptions, narratives, situations, or other questions that you felt were in any way 'unfair'?"

As shown in Figures 3 and 4, virtually no students felt that they had been asked any "unfair" questions by the live interviewer, while 30 percent felt that at least some portion of the taped test had included such questions. Write-in comments indicated that, for the most part, students were referring in this item to particular questions they had not been able to deal with properly for lack of proficiency, rather than as a result of intrinsic flaws in the test questions or testing procedures per se. However, two students suggested that the directions for the series-of-pictures section should be revised to indicate

Figure 1

ABILITY PROBED IN LIVE INTERVIEW?

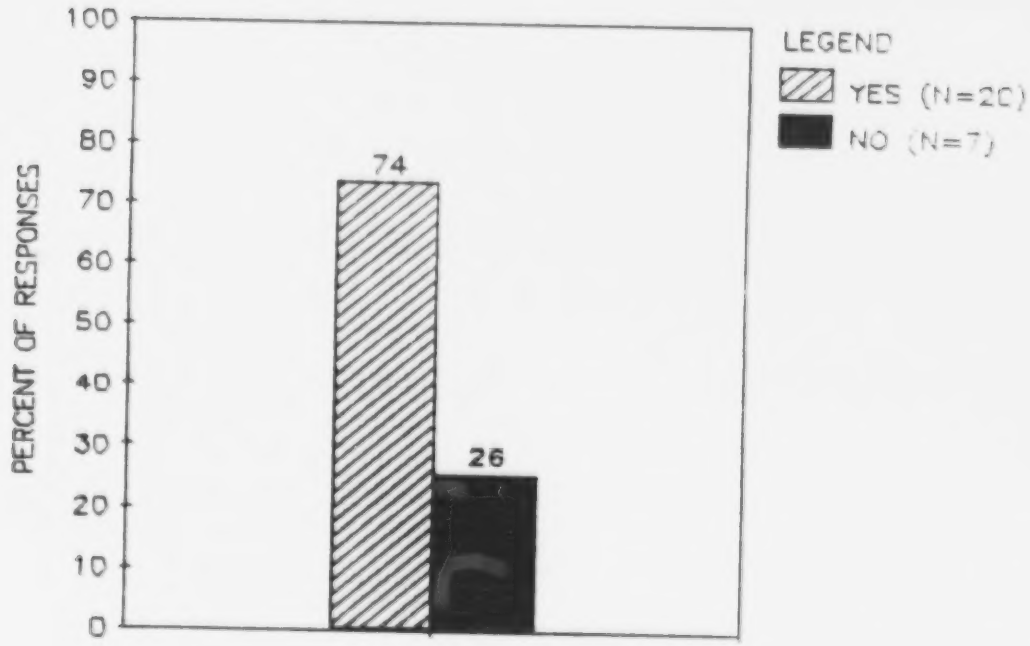


Figure 2

ABILITY PROBED IN TAPED TEST?

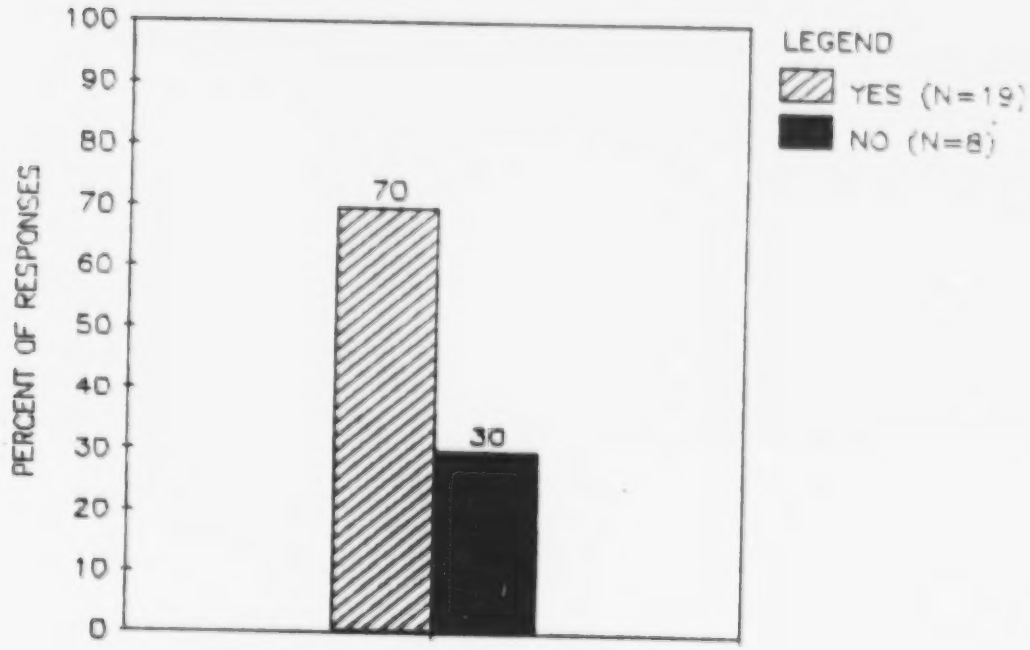


Figure 3

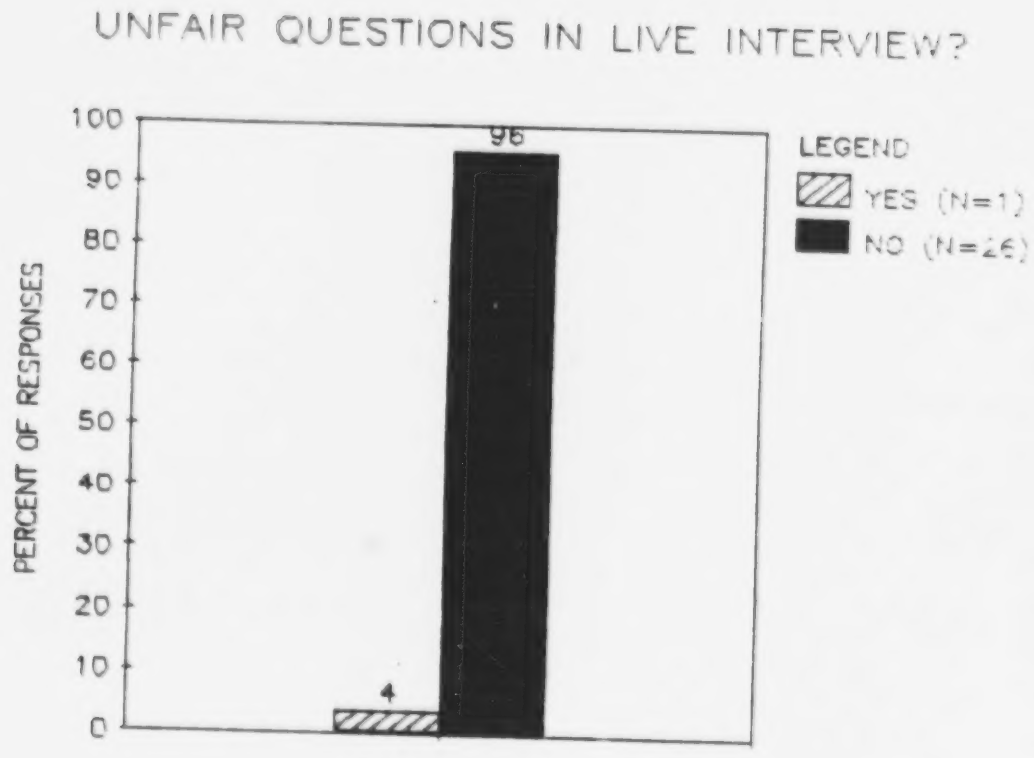


Figure 4

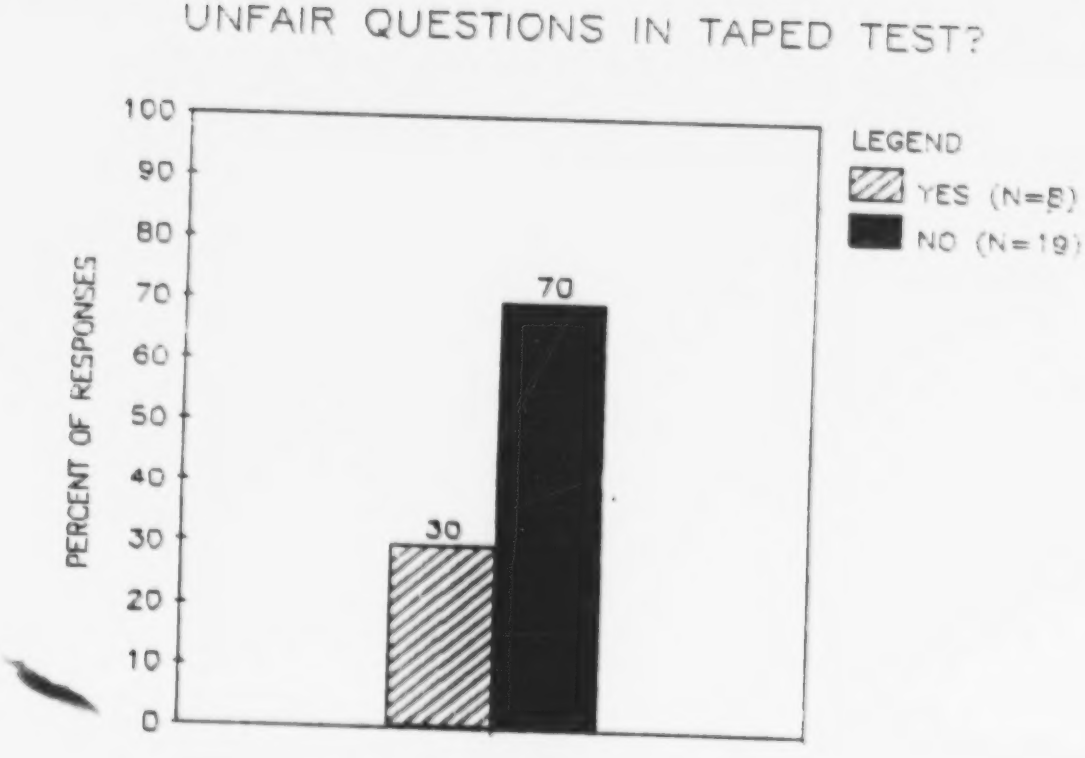


Figure 5

IN WHICH TEST MORE NERVOUS?

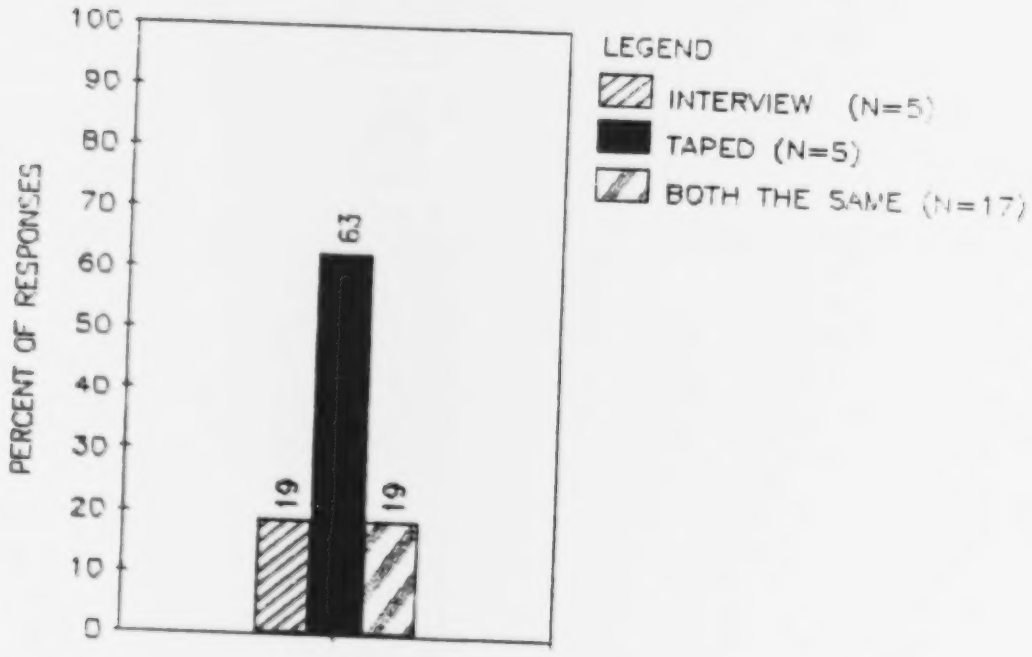


Figure 6

WHICH TEST MORE DIFFICULT?

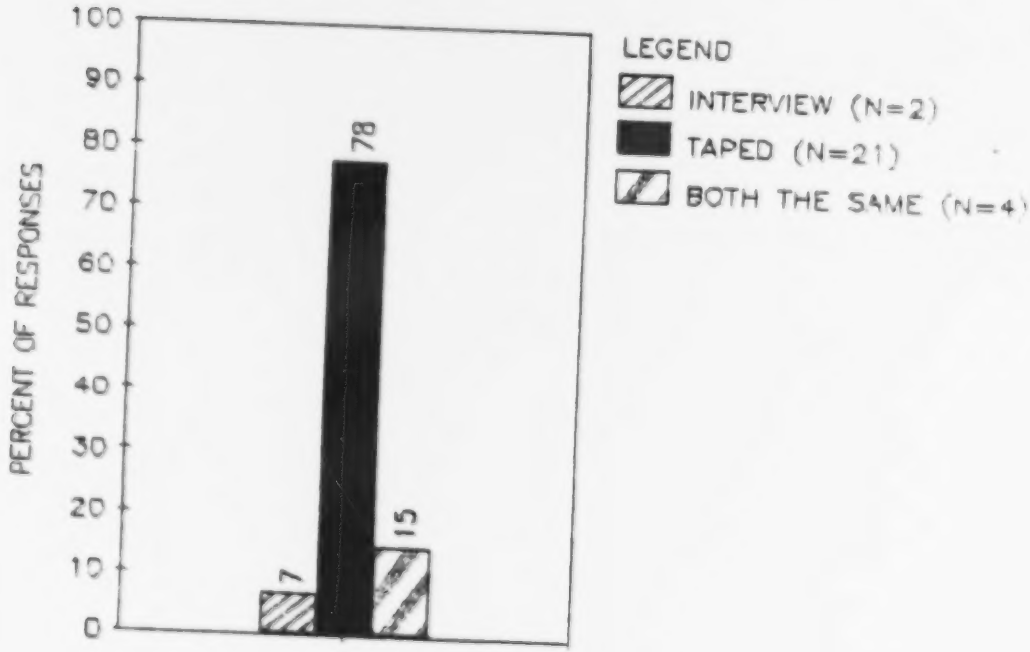


Figure 7

TAPE PAUSES LONG ENOUGH?

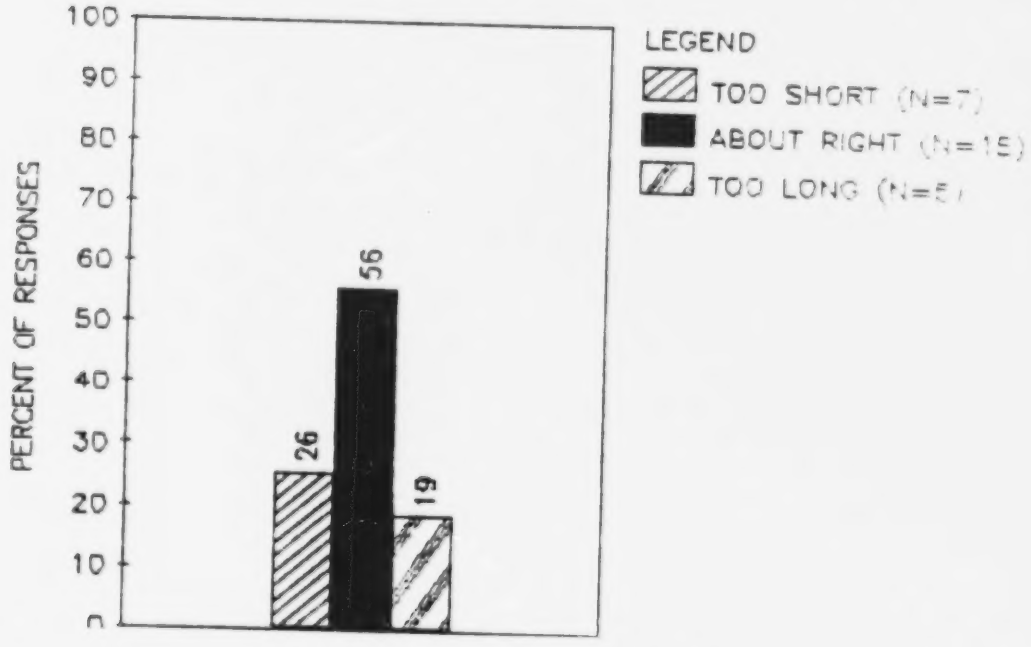


Figure 8

TEST DIRECTIONS CLEAR?

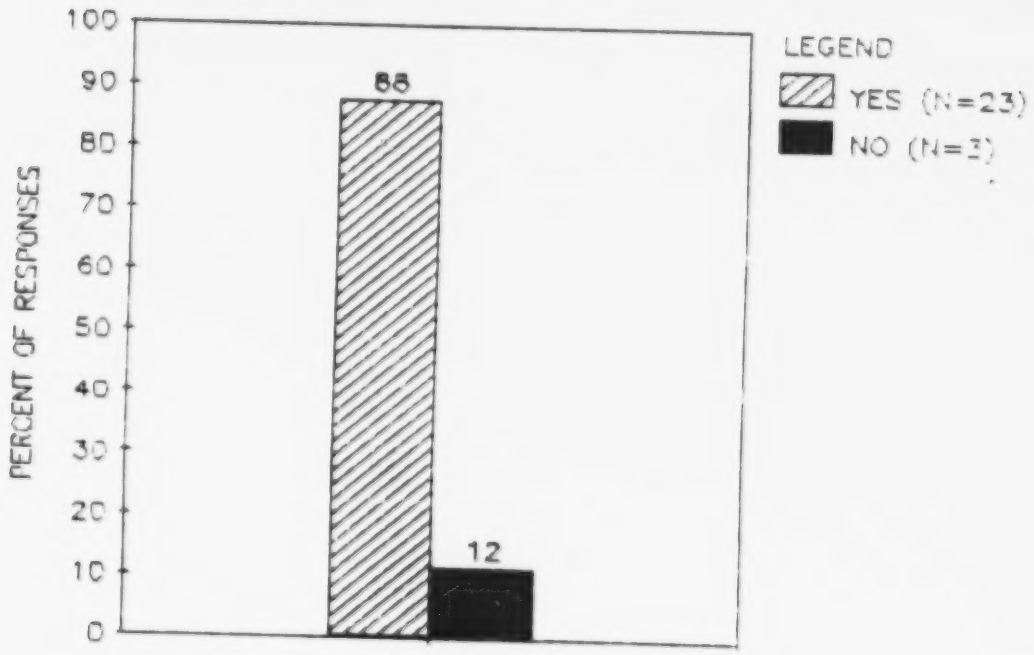
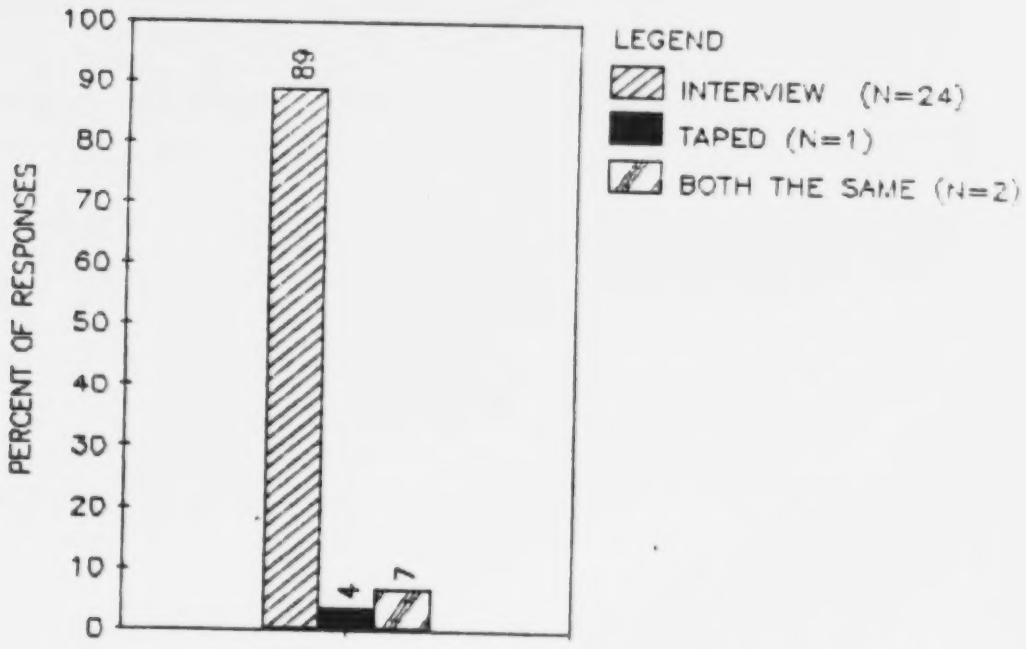


Figure 9

PREFER LIVE INTERVIEW OR TAPED?



more clearly that each of the pictures in the set should be addressed in a sequential manner.

Figure 5 summarizes the responses to the question, "In which of the two types of test--live interview or taped test--did you feel the more anxious or nervous?" A clear majority (63%) of respondents felt that they had been more nervous during the taped test, while equal numbers (19% in each case) were divided between considering both types of test equally nervousness-producing or attributing this characteristic predominantly to the live interview.

Notwithstanding the essentially equivalent scores which they obtained under both the live and taped tests (scoring results were not communicated to the students until several days after questionnaire administration), the great majority (78%) considered the taped test "more difficult" than the live interview, with only 7 percent having the opposite opinion (Figure 6).

With regard to certain technical aspects of the taped test, most of the respondents (56%) felt that the length of pauses provided on the tape was "usually about right" for them to respond as fully as they desired or were able. Pauses were considered generally "too long" by 19 percent and "too short" by 26 percent (Figure 7). A large majority (85%) considered the taped test directions "sufficiently clear and detailed," with only 12 percent of the contrary opinion (Figure 8).

To the "bottom-line" question, "Assuming that you would receive the same score through both techniques, would you personally rather take a live interview or a (single) taped test in order to show your speaking proficiency?," examinee responses were overwhelmingly (89%) in favor of the live interview, with only 4 percent expressing a preference for the taped test (Figure 9).

The overall results of this brief survey of student opinions concerning the semi-direct testing procedure, both in its own right and by comparision to direct face-to-face interviewing, appear to suggest that while students view the taped test as generally well constructed and sufficiently probing from the standpoint of elicitation procedures, they feel it is more difficult than the live interview and tend to consider at least portions of the test as "unfair." In a forced choice between the two types of testing, the great majority of examinees indicate a personal preference for undergoing a live interview rather than a tape recorded test. From an administrative viewpoint, implications of the student questionnaire data would seem to be that face-to-face interviewing is preferable whenever the necessary resources can be made available, but that when an alternative approach is required, the students involved will generally consider themselves adequately tested through semi-direct means, albeit as a "second choice" procedure.

DISSEMINATION OF STUDY RESULTS

Camera-ready copy for all four forms of the tests developed under this project, as well as master stimulus tapes, is presently housed at the Center for Applied Linguistics (CAL). CAL intends to make copies of the test materials, as well as a test scoring service, available to the field on a cost-recovery fee basis, within the near future. Copies of the test development handbook will also be available through CAL.

References:

Clark, John L. D. "Direct vs. Semi-Direct Tests of Speaking Proficiency," pp. 35-49 in Eugène J. Brière and Frances B. Hinofotis, eds., Concepts in Language Testing: Some Recent Studies. Washington, DC: Teachers of English to Speakers of Other Languages, 1979.

Lowe, Pardee, Jr. and Ray T. Clifford. "Developing an Indirect Measure of Overall Oral Proficiency," pp.31-39 in James R. Frith, ed., Measuring Spoken Language Proficiency. Washington, DC: Georgetown University Press, 1980.